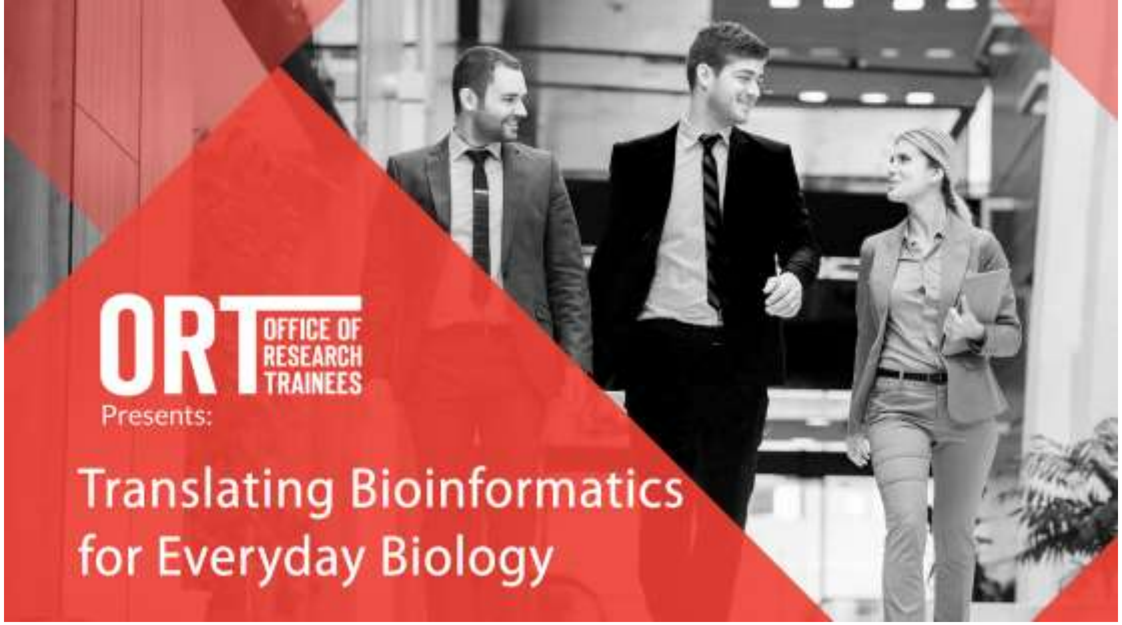


RNA-seq analysis

Musa Ahmed

Jan 31st, 2017

GOAL

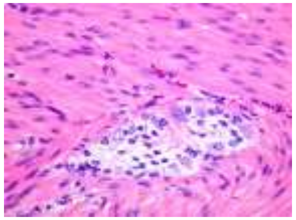
- Basics of RNA-seq analysis
- Applications
- Challenges
- Practical
  - Alignment
  - DGE analysis

What is RNA-seq

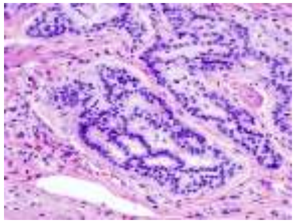
- RNA-seq works by sequencing every RNA molecule and profiling the expression of a particular gene by counting the number of time its transcripts have been sequenced.

RNA-seq

Samples of interest

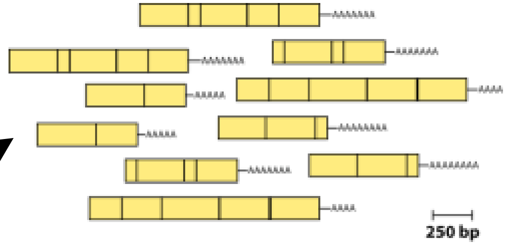


Condition 1
(normal colon)

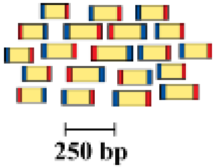


Condition 2
(colon tumor)

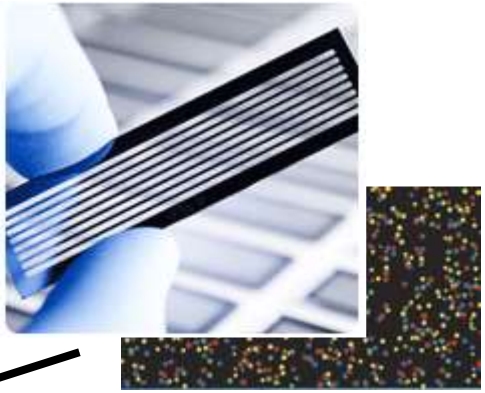
Isolate RNAs

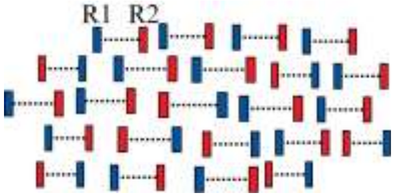


Generate cDNA, fragment, size select, add linkers

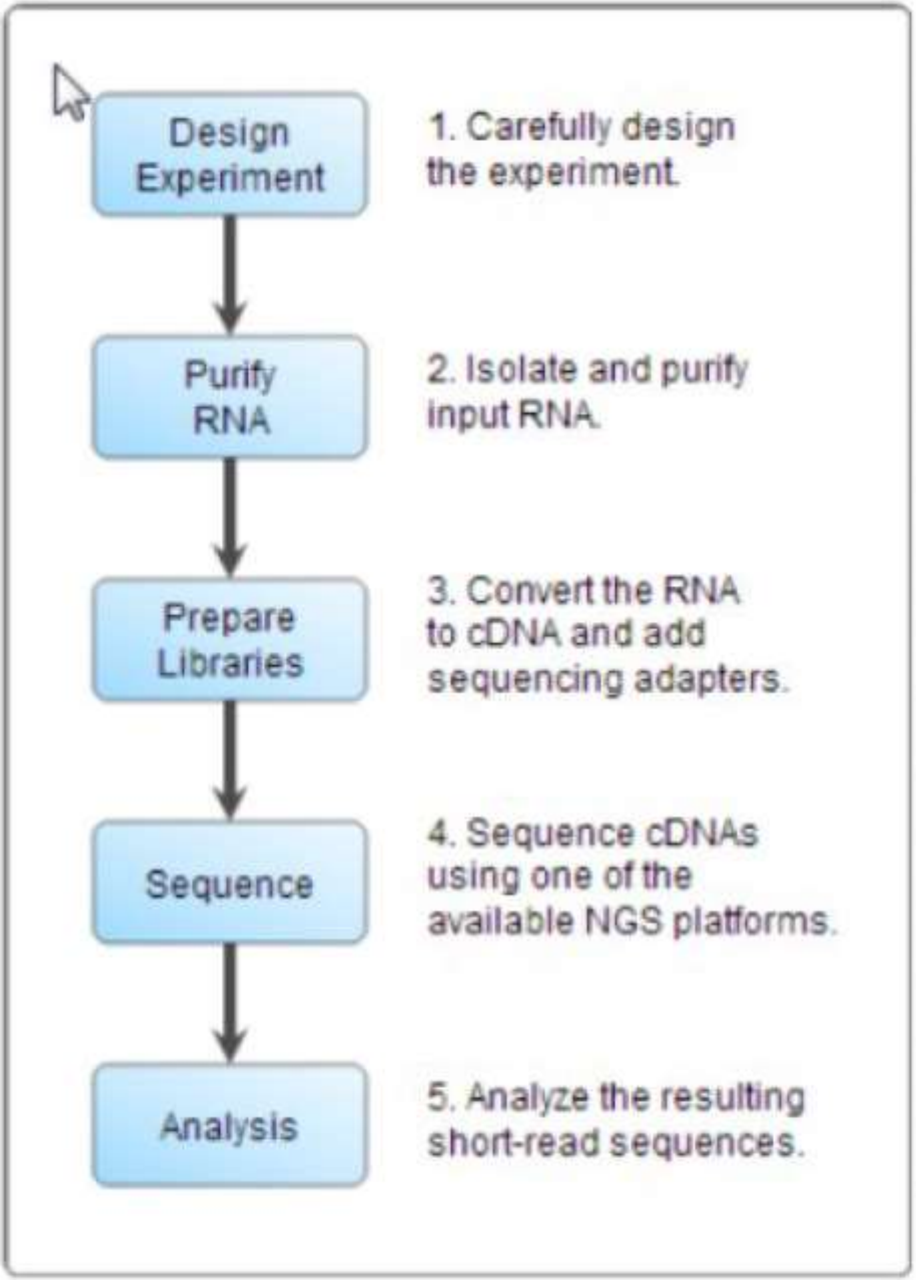


Sequence ends





100s of millions of paired reads
10s of billions bases of sequence



Design Experiment

1. Carefully design the experiment.

Purify RNA

2. Isolate and purify input RNA.

Prepare Libraries

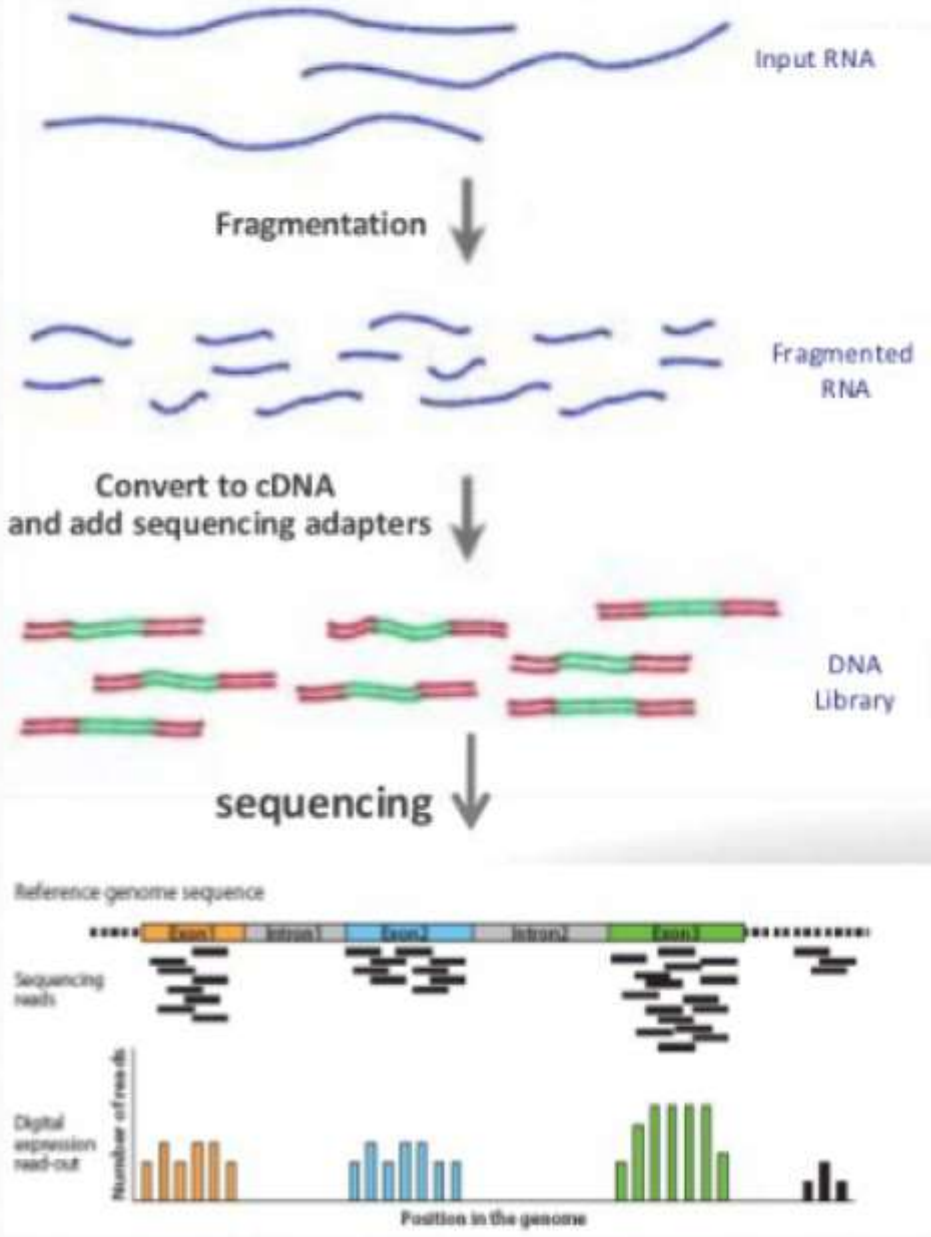
3. Convert the RNA to cDNA and add sequencing adapters.

Sequence

4. Sequence cDNAs using one of the available NGS platforms.

Analysis

5. Analyze the resulting short-read sequences.

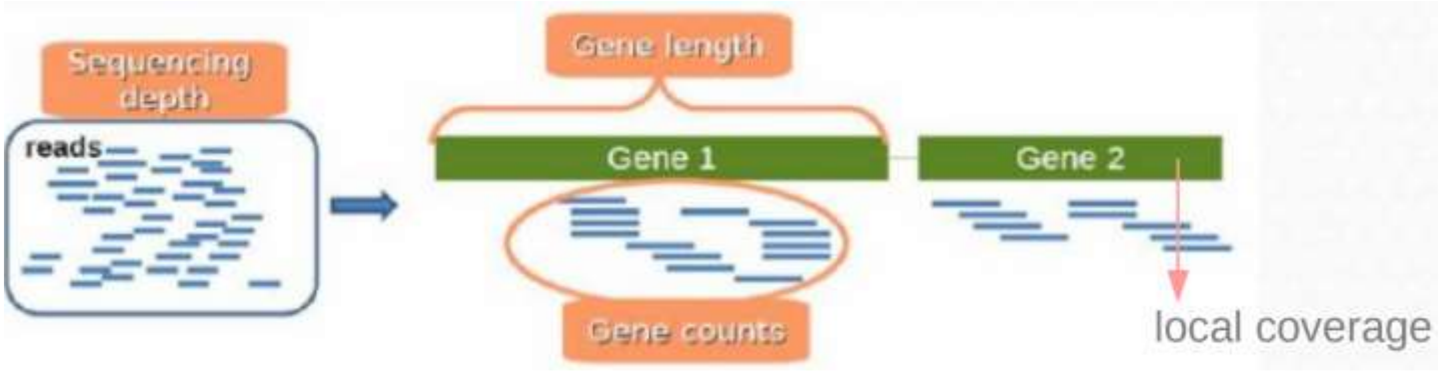


Applications

- Gene expression and differential expression
- Transcript discovery
- SNV, RNA-editing events, variant validation
- Allele specific expression
- Gene fusion events detection
- Genome annotation and assembly
- etc …

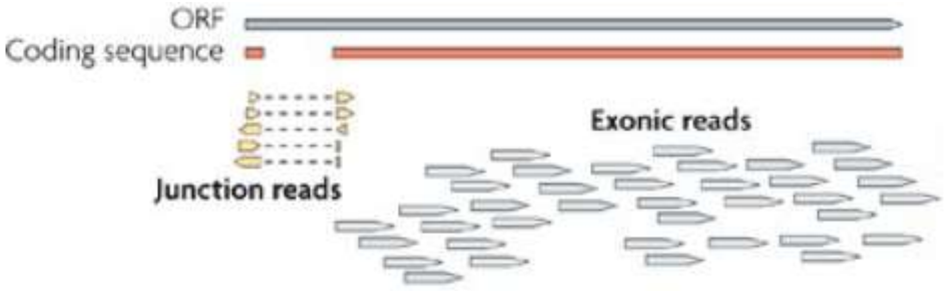
Key concepts

- **Sequencing depth**
  - Total number of reads mapped to the genome. (Library size) Could also be applied to samples.
- **Coverage**
  - Number of reads mapped to a specific region (average of them if we are talking about the whole genome...)
- **Gene length**
  - Number of bases that a gene has.

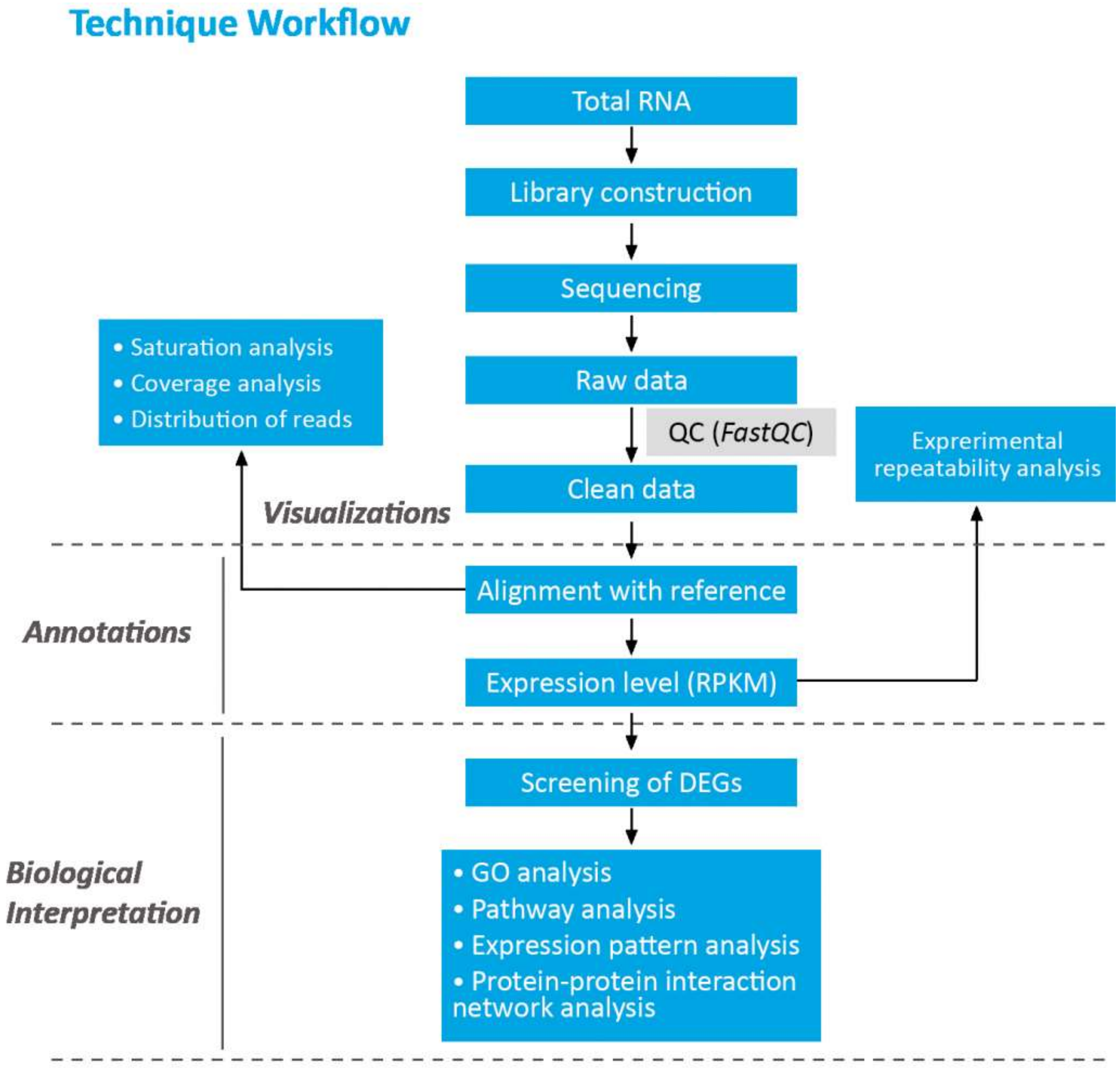


Key concepts

- **Exonic reads:** Reads within exons
- **Junction reads:** Reads spanning exon junctions



BGI workflow



Technique Workflow

Total RNA

Library construction

Sequencing

Raw data

QC (*FastQC*)

Clean data

- Saturation analysis
- Coverage analysis
- Distribution of reads

Visualizations

Exprerimental repeatability analysis

Annotations

Alignment with reference

Expression level (RPKM)

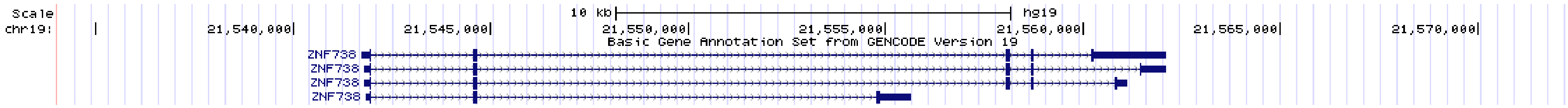
Biological Interpretation

Screening of DEGs

- GO analysis
- Pathway analysis
- Expression pattern analysis
- Protein-protein interaction network analysis

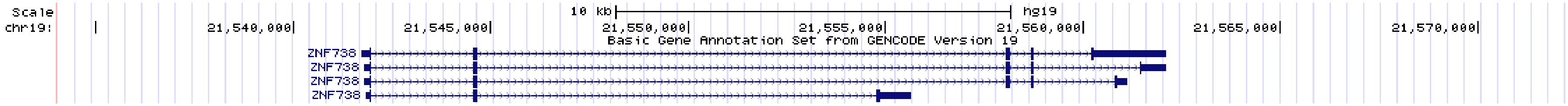
Reference genomes by Genome Reference Consortium

- Assembled whole genome using a bunch of individual genome sequences
  - Human – Hg38 (GRCh38), hg19 (GRCh37), b37 etc
  - Mouse – mm10, mm9 etc



Gene annotations

- Annotation of the whole genome (protein coding genes, noncoding RNAs etc)
    - RefSeq (good for general analysis)
    - GENCODE/ENSEMBL (good for noncoding genes)
    - miTranscriptome (good for noncoding genes)



Annotations can be downloaded from UCSC genome browser

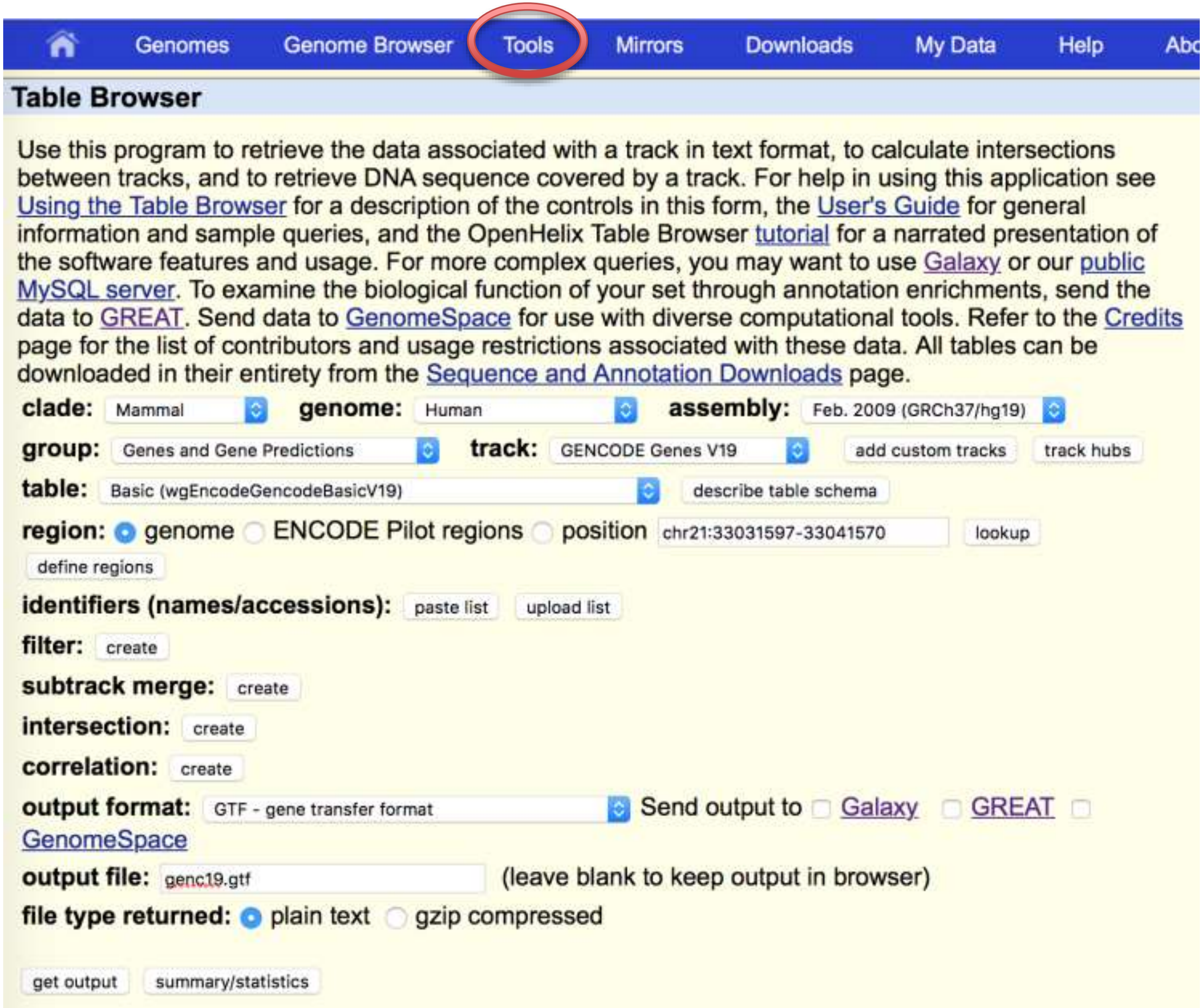


Table Browser

Use this program to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA sequence covered by a track. For help in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for general information and sample queries, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more complex queries, you may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the data to GREAT. Send data to GenomeSpace for use with diverse computational tools. Refer to the Credits page for the list of contributors and usage restrictions associated with these data. All tables can be downloaded in their entirety from the Sequence and Annotation Downloads page.

clade: Mammal genome: Human assembly: Feb. 2009 (GRCh37/hg19)

group: Genes and Gene Predictions track: GENCODE Genes V19 add custom tracks track hubs

table: Basic (wgEncodeGencodeBasicV19) describe table schema

region: ● genome ○ ENCODE Pilot regions ○ position chr21:33031597-33041570 lookup

define regions

identifiers (names/accessions): paste list upload list

filter: create

subtrack merge: create

intersection: create

correlation: create

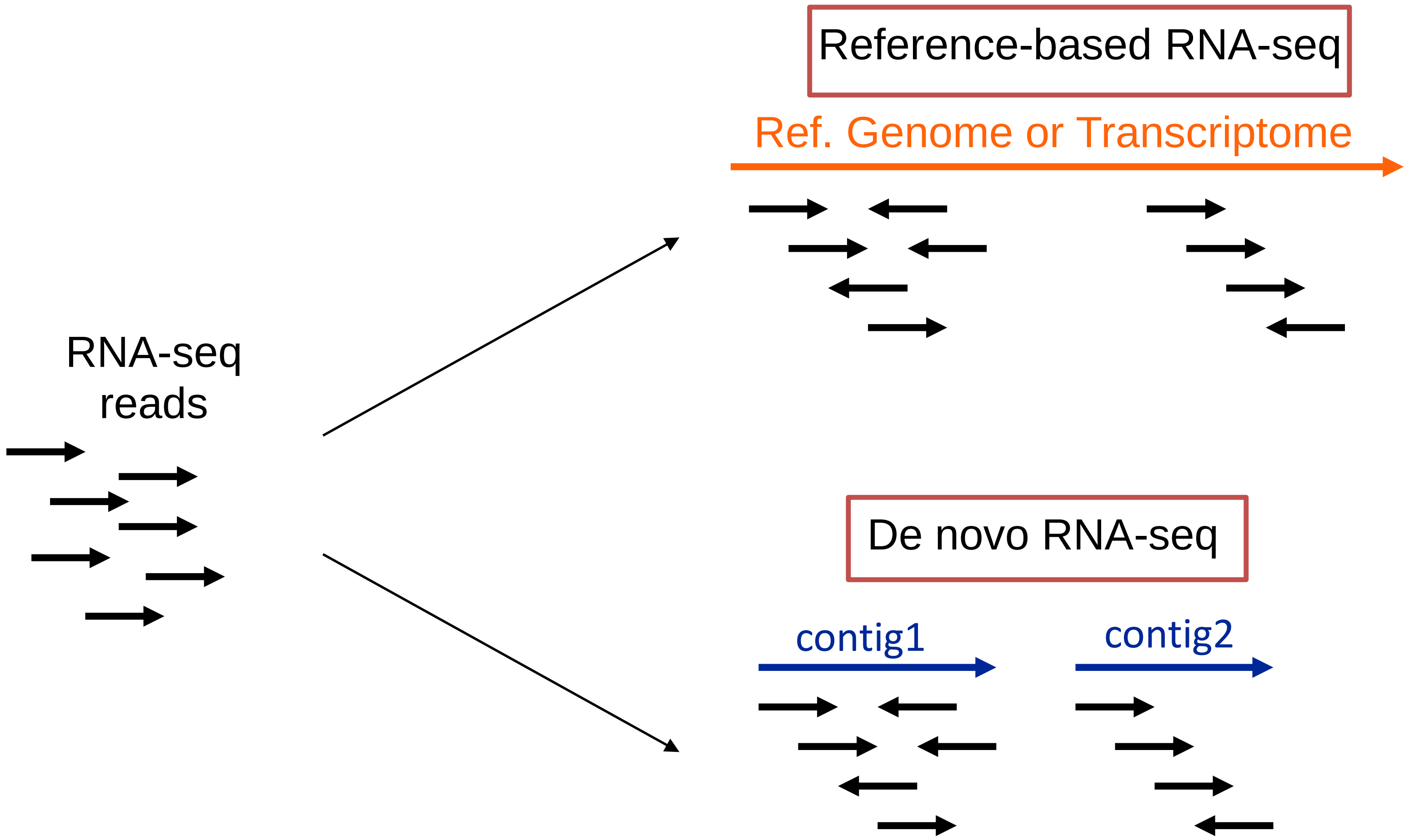
output format: GTF - gene transfer format Send output to □ Galaxy □ GREAT □ GenomeSpace

output file: genc19.gtf (leave blank to keep output in browser)

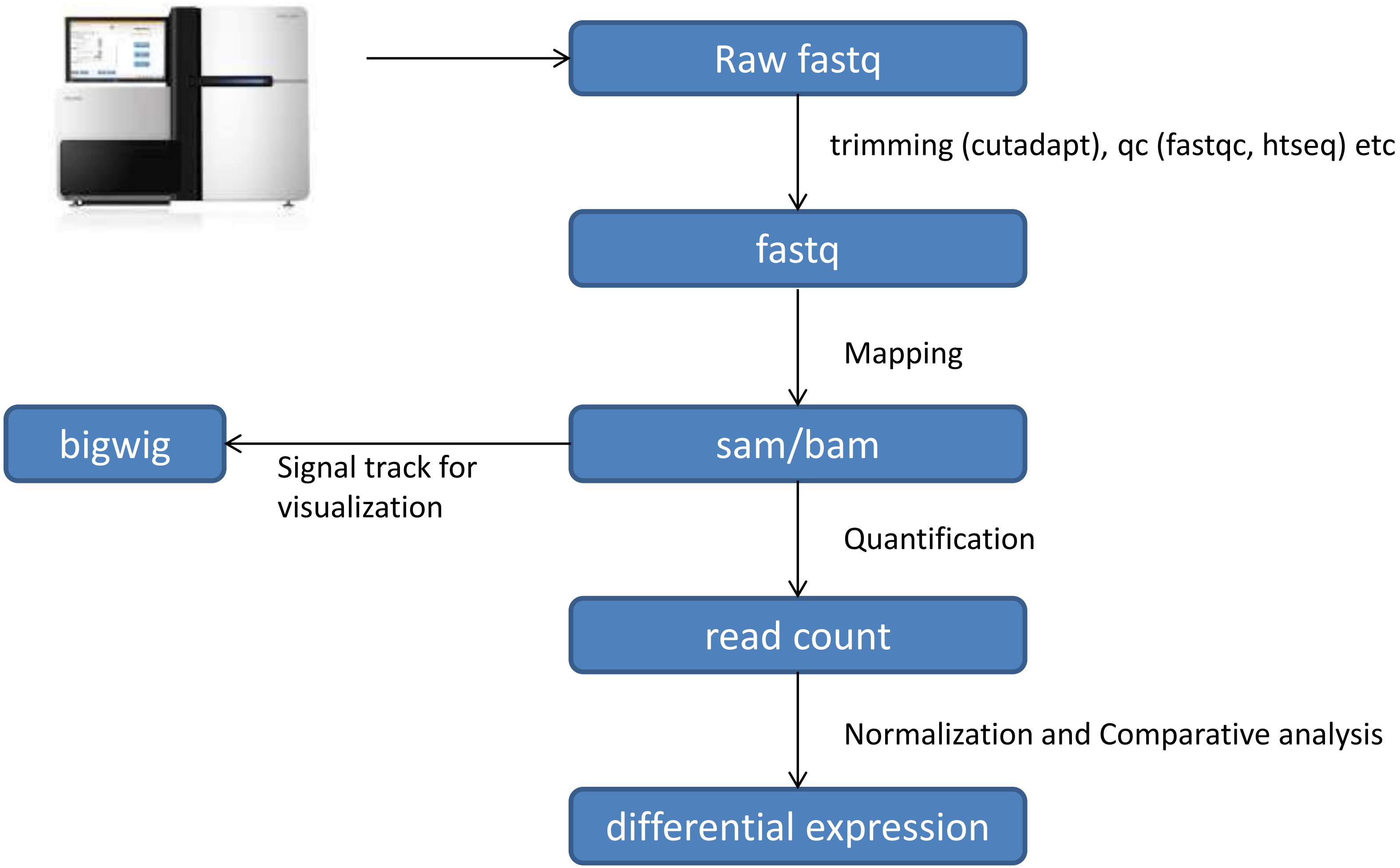
file type returned: ● plain text ○ gzip compressed

get output summary/statistics

RNA-seq alignment can be annotation-dependent or de novo



RNA-seq analysis: overview



Raw fastq

trimming (cutadapt), qc (fastqc, htseq) etc

fastq

Mapping

sam/bam

bigwig

Signal track for visualization

Quantification

read count

Normalization and Comparative analysis

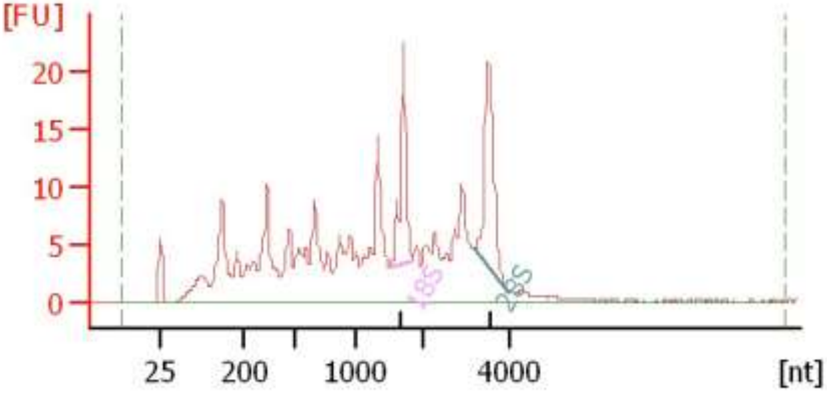
differential expression

RNA-seq questions during library preparation

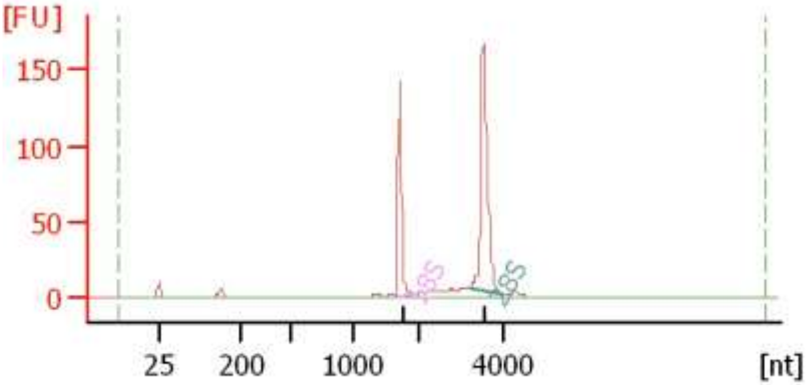
- Construction strategies
  - total RNA or polyA RNA?
  - Ribo minus or plus?
  - Stranded or unstranded?
  - size – microRNA?
- RNA quantity
- RNA quality
  - RNA is fragile and easily degraded
  - Low quality material can bias the data
- Replicates

Agilent

- https://github.com/griffithlab/rnaseq_tutorial/wiki/Resources/Agilent_Trace_Examples.pdf
- 'RIN' = RNA integrity number
  - 0 (bad) to 10 (good)



RIN = 6.0



RIN = 10

RNA-seq questions during mapping

- Reference genome version – the latest version may have compatibility issues with other analysis
- Annotation – refseq or gencode or ENSEMBL
- Want junction read or not
- Remove duplicates? (No!)
- How many missmatches to allow?

RNA-seq questions during quantification

- Keep reads mapping to multiple loci?
- Keep reads overlapping multiple genes?

Alignment/Mapping tools

- TopHat2 (widely used, slow)
- STAR (super fast, very popular now, very demanding)
- RSEM (getting popular, used by ENCODE)
- Rsubread (works in R, not very popular)
- Sailfish (good for isoforms, less popular)

Quantification

- TopHat2 → Cufflinks2 (provides FPKM) → Cuffdiff (DGE analysis)

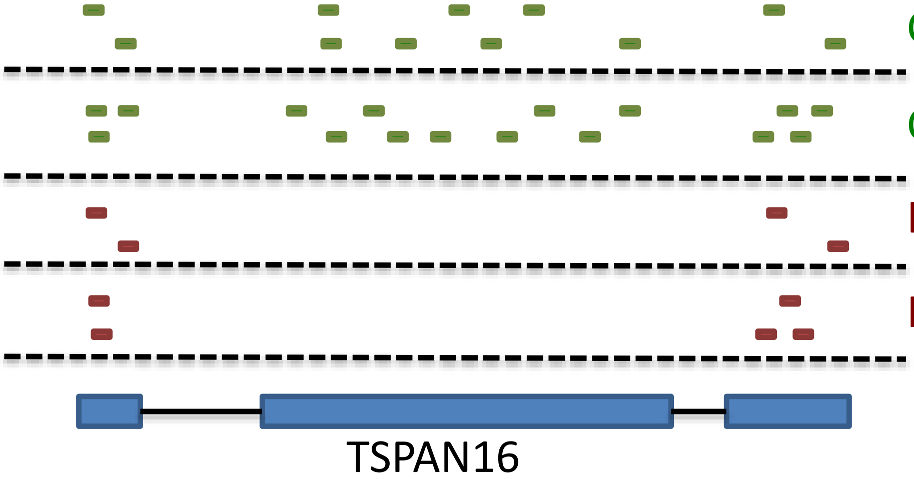
- STAR → HTSeq-count, featureCount (provides raw count) → DESeq2, EdgeR (DGE analysis and normalized count)

- RSEM → RSEM (provides expected count, TPM and FPKM) → EBSeq

- Rsubread → featureCount → DESeq2, EdgeR

- Sailfish →Sailfish (provides raw count, TPM) → DESeq2, EdgeR

Summarized RNA-seq



Contro1: **11** reads

Control2: **16** reads

KnockDown1: **4** reads

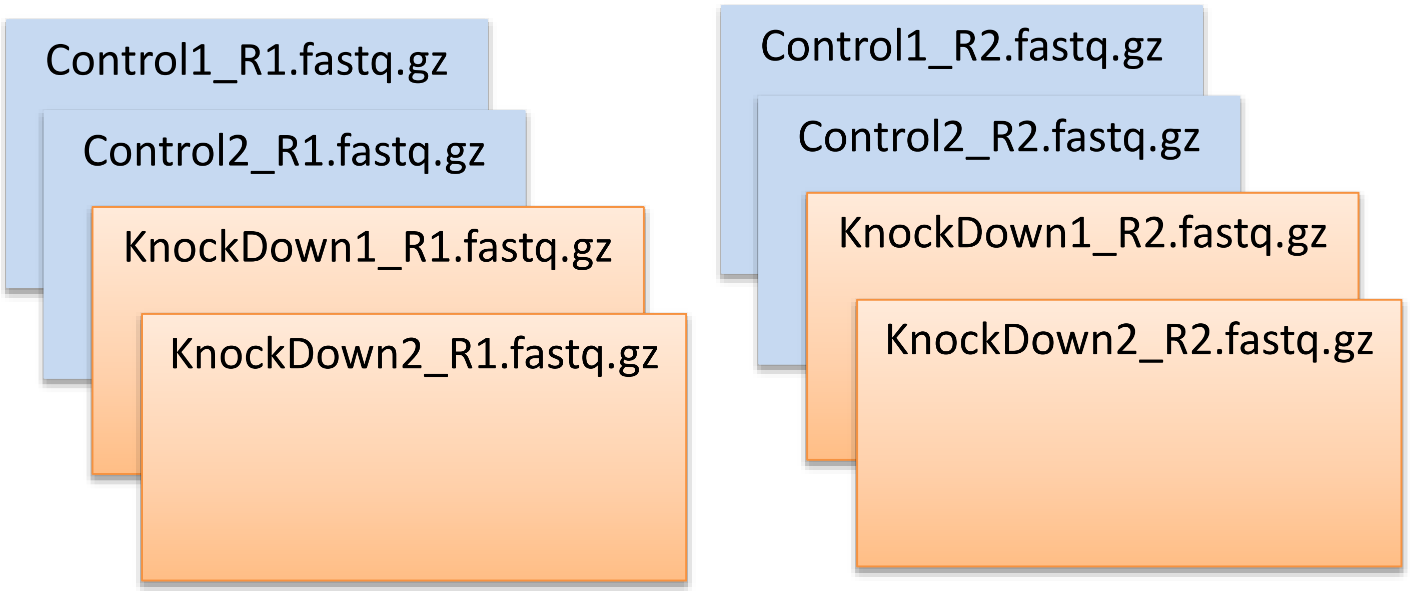
KnockDown2: **5** reads

	Control1	Control2	KnockDown1	KnockDown2
TSPAN6	11	16	4	5
TNMD	1	0	0	0
DPM1	435	743	836	739
SCYL3	203	218	416	352
C1orf112	216	643	714	704
FGR	2365	5011	2828	2294
CFH	6	1	4	0
FUCA2	380	865	431	523
...	...	...	...	...
NFYA	888	827	1674	1580

Key concepts

- **Expression units:** There are several expression units available – RPKM/FPKM, CPM, TPM, Normalized expression
- **Within sample normalization:** Expression normalized between genes within the sample (e.g., FPKM)
- **Between sample normalization:** Expression normalized between samples (e.g., TPM)
- **Fasta file:** Sequence storing file (can be opened in TextWrangler (unix) or Notepad++ (windows))
  - Format:
    ```
    >sequence1
    ATCGTGCTGATGCGTGACG
    ```
- **Fastq file:** Sequence storing file with quality score, what you get from the sequencing centre

First file: fastq



Control1_R1.fastq.gz

Control2_R1.fastq.gz

KnockDown1_R1.fastq.gz

KnockDown2_R1.fastq.gz

Control1_R2.fastq.gz

Control2_R2.fastq.gz

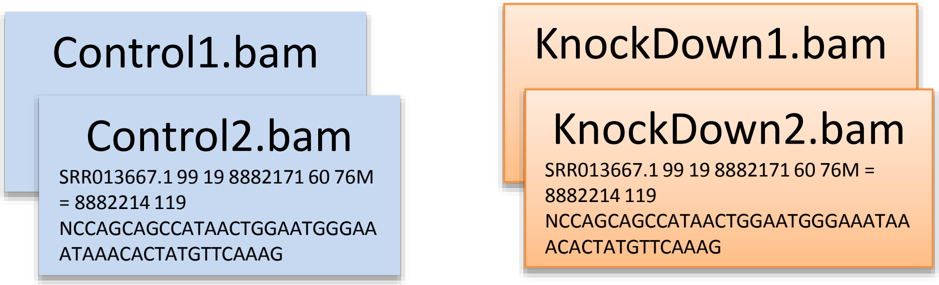
KnockDown1_R2.fastq.gz

KnockDown2_R2.fastq.gz

~ 10Gb each sample

```
@ERR127302.1 HWI-EAS350_0441:1:1:1055:4898#0/1
GGCTCATCTTGAACTGGGTGGCGACCGTCCCTGGCCCCTTCTTGACACCCA
+
4=B@D99BDDDDDDD:DD?B<<=?>6B########################
```

From fastq to sam/bam



Control1.bam

Control2.bam

SRR013667.1 99 19 8882171 60 76M = 8882214 119 NCCAGCAGCCATAACTGGAATGGGAAATAAACACTATGTTCAAAG

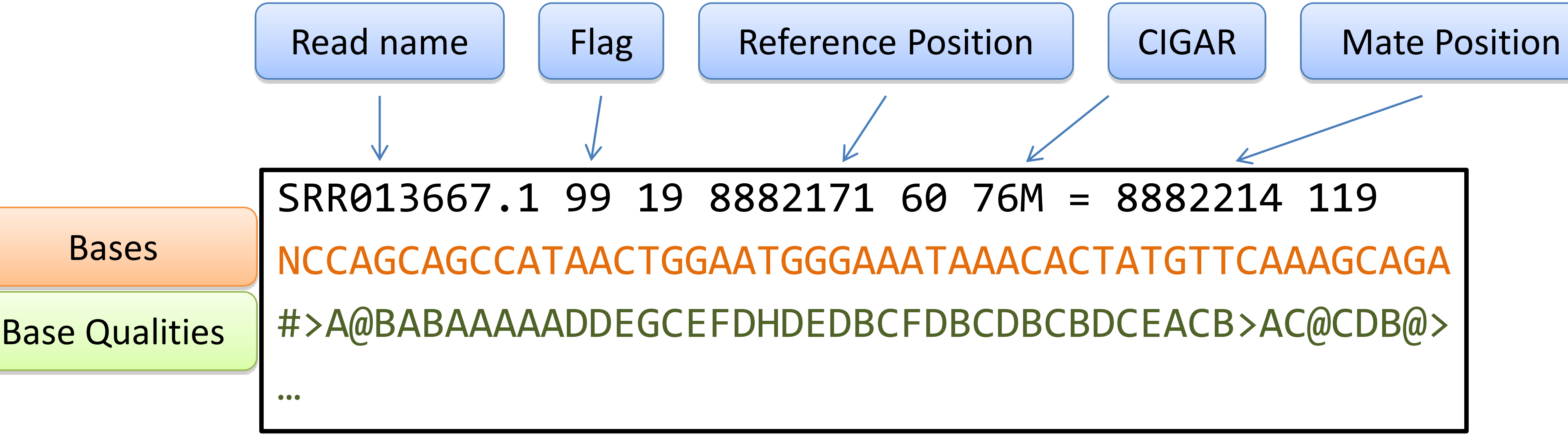
KnockDown1.bam

KnockDown2.bam

SRR013667.1 99 19 8882171 60 76M = 8882214 119 NCCAGCAGCCATAACTGGAATGGGAAATAAACACTATGTTCAAAG

~ 10Gb each bam

- Used to store alignments
- SAM = text, BAM = binary



Read name

Flag

Reference Position

CIGAR

Mate Position

Bases

Base Qualities

```
SRR013667.1 99 19 8882171 60 76M = 8882214 119
NCCAGCAGCCATAACTGGAATGGGAAATAAACACTATGTTCAAAGCAGA
#>A@BABAAAAADDEGCEFDHDEDBCFDBCDBCBDCEACB>AC@CDB@>
…
```

SAM: Sequence Alignment/Map format

Trivia time!

- What is the first step after getting the fastq file?
  - a) Forward to bioinformatician b) Alignment c) Ignore that it's there d) Quantification
- Should we always have replicates?
  - a) Yes b) No
- From fastq we make bam files. What do bam files contain?
  - a) Mapped reads b) Treasure map c) Raw reads d) Nothing useful really
- Can I use FPKM in DESeq2?
  - a) Yes b) No
- Which one of below is a major bottleneck in gene expression analysis?
  - a) My will to do it b) High performance computers c) Repeats in the genome d) Donald Trump
- What data should I use to generate an expression boxplot for *MYC* for 4 samples processed together?
  - a) raw read count b) FPKM c) normalized read count d) TPM
- HTSeq-count or featureCount requires ________
  - a) fastq files b) love c) bam files d) Donald Trump

Resources

- http://www.bioconductor.org/help/workflows/rnaseqGene
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4728800/
- https://usegalaxy.org/u/jeremy/p/galaxy-rna-seq-analysis-exercise